Journal of Consulting and Clinical Psychology
1970, Vol. 35, No. 2, 143–147

FROM THE SUBJECT'S POINT OF VIEW, WHEN IS BEHAVIOR PRIVATE AND WHEN IS IT PUBLIC:

PROBLEMS OF INFERENCE [1]

MARTIN T. ORNE [2]

Institute of the Pennsylvania Hospital and University of Pennsylvania

In an invited discussion, methodological questions concerning McFall's study of the effect of self-monitoring on smoking behavior are raised. It is emphasized that results of such studies should be evaluated from the point of view of the *S* rather than the investigator. Some measures assumed to be unobtrusive by the investigator share qualities of deception experiments: It must be determined, therefore, whether it is the *S* or the *E* who is deceived. Procedures that may be helpful in clarifying such questions and the difficulties of generalizing results to other contexts are discussed.

A number of very significant and topical issues are raised by McFall (1970). The problem of criterion measures has long bedeviled all psychotherapy research. Behaviorally oriented therapists have generally been in a better position since they were able to specify the behavior to be modified. As McFall's study clearly demonstrates, some of the measures commonly utilized in evaluating the effects of behavioral interventions are in themselves altered by the very act of collecting data. Because self-reports that require an *S* to monitor his own behavior are easily obtained and are generally very sensitive to change, the technique has been widely utilized. However, if requiring an *S* to report his own behavior in and of itself modifies the behavior, care must be taken in accepting at face value findings based on such evidence.

In an elegant way McFall reasons that if asking an *S* to monitor his behavior will have an effect on the behavior under surveillance, it should make a difference what the *S* is asked to monitor. He illustrates this general point by translation into a timely and important paradigm: It is argued that requiring an individual to report the *number of cigarettes smoked* may have an aversive conditioning effect, and asking him instead to indicate each time he successfully *resists the temptation to smoke* should therefore reduce cigarette consumption. Empirically, his study demonstrates not only that asking *S*s to note each time they smoke a cigarette will affect their actual smoking behavior, but he also shows a clear-cut difference in the amount smoked, depending on whether *S*s were asked to record the frequency of the impulse to smoke or the number of cigarettes consumed. This finding—potentially of very great practical as well as theoretical importance—well deserves further investigation.

While the conclusions reached seem intuitively compelling, it is unclear to me whether, from the *S*'s point of view in this experiment, the behavior may have other alternative and instructive meaning. Because the present study not only deals with an extremely interesting problem, but also was carefully carried out and executed, it serves to illustrate major differences in the interpretation of the data between analyzing the experiment from the viewpoint of the *S* rather than the *O*. These views would be obscured in a piece of research that was inferior and marred by obvious methodological flaws. However, because of basic issues of interpretation, of how the experiment might have been perceived by the *S*, I suspect that some of the findings in the study are unique to the particular experimental situation employed and will not generalize outside of the specific experimental setting. It should be clear that my comments are intended not as a critique of an interesting and worthwhile investigation, but rather to raise issues for future research. In the analysis to follow, I will make assumptions about what I believe are plausible alternative explanations of the author's findings. The data to support these assumptions are not for the most part available; nonetheless, it seems

[1] The substantive methodological studies on which this discussion is based were originally supported in part by the Office of Naval Research, Contract No. Nonr 4731(00). The author wishes to express his appreciation for the helpful comments of his colleagues, Frederick J. Evans, Kenneth R. Graham, Emily C. Orne, David A. Paskewitz, and Harvey D. Cohen.

[2] Requests for reprints should be sent to Martin T. Orne, Unit for Experimental Psychiatry, Pennsylvania Hospital, 111 North Forty-Ninth Street, Philadelphia, Pennsylvania 19139.

fruitful to state these possibilities in a positive fashion which will lend itself to empirical testing.

The observation that asking a patient to monitor his own behavior has consequences for the behavior is generally well recognized by clinicians. A particularly good example is the common observation of physicians working with obesity—when patients can be induced to keep an accurate diary of their food intake for the purpose of "establishing a base line," a loss of weight is usually observed (A. J. Stunkard, personal communication, 1969). It would seem that the patient who is constrained by the procedure to admit his actual food intake both to himself and the therapist, tends to forego some of the items he would otherwise have consumed. Unfortunately it is difficult to persuade patients to keep an honest record of their food intake.

The reactive aspects of keeping a food diary can be evaluated relatively easily because the patient's weight serves as a convenient and reliable measure of total caloric intake. The loss of weight, therefore, reflects a decrease in eating behavior. Unfortunately, no equally reliable way is available to test the individual's overall cigarette consumption, yet such an overall measure would be of great importance in evaluating procedures designed to curtail smoking. Therefore, McFall had to develop analogous information in a more complex fashion.

In order to draw inference about the effect of instructions on smoking behavior in general, it was necessary to utilize detailed observations of this behavior during a specified period where observational data could be obtained. The author uses the number of cigarettes smoked during classroom hours as an index of overall smoking behavior. Unfortunately classroom smoking is not necessarily a representative sample of smoking behavior in general. Some individuals tend to smoke primarily in situations of emotional stress, others when they are working, others when they are bored, still others in social situations when they need a prop, etc. The effectiveness of instructions in modifying overall smoking behavior would not necessarily be reflected adequately by smoking that could be observed in the classroom. Beyond this general problem, however, there are more serious difficulties in generalizing from such data.

The study was carried out using the investigator's own students in a psychology class. The nonsmokers, as determined by a questionnaire, were asked to participate as *O*s to keep track of the behavior of the smoking individuals who would later be the *S*s of the experiment. In order to do this, all students were required to fill in a questionnaire about many aspects of their behavior when the class first met; nonsmokers were then segregated (presumably on a totally unrelated basis), asked to stay after class, and their cooperation as collaborators was elicited and obtained. These *O*s were each assigned to observe one smoking *S* seated close to them; it was their task to keep track of the number of cigarettes smoked—initially during a two-week period of actual base line prior to any mention of smoking behavior during the course, and later after the smoking members of the class had been asked to monitor their own smoking behavior and report it. At that time, half of the smokers were asked to monitor the number of cigarettes smoked, and the other half, the number of times they felt like smoking but did not.

The purpose of eliciting cooperation from the nonsmoking *O*s is to test the reactivity of self-monitoring by the smoking *S*s. It is, of course, essential that these individuals be unaware that their smoking behavior is being monitored, either during their base line or subsequently. In many ways, then, this experiment has the attributes of a deception study, in that it depends on the *O*'s keeping accurate track of his target *S*'s smoking behavior without letting *S* know he is being observed. While only one smoking *S* admitted that his nonsmoking *O* had told him about the procedure, it seems unlikely that this was limited to only one *S*. As *S*s did not volunteer for their task, but rather were drafted into both the role of *O* and *S*, the likelihood that the investigator's interest in smoking behavior was known on campus, and given the probability that *S*s had contact with each other outside of class, it seems plausible that what I have described as a "pact of ignorance" (Orne, 1962) could have existed between the investigator and several of his *S*s. The fact that it was necessary for the investigator to model smoking in order to obtain a sufficiently high base rate increased the odds that the smoking *S* recognized his role in the experiment. As in any study involving deception, it is essential to determine the extent to which it is the *S* or the *E* who is deceived. (For a discussion of these issues, see Orne & Holland, 1968.)

Among the findings reported, one of the more interesting is the discrepancy between the number of cigarettes reported smoked by the *S*s, as compared with those reported by the *O*s of these *S*s. Typically, *O*s reported fewer cigarettes smoked than did the *S*s themselves who were instructed to report their smoking behavior. No self-report data about smoking are available for those *S*s who had been instructed to report only the impulse to smoke. However, smoking a cigarette

takes a certain period of time, and it is difficult to see how an *O*, given the relatively easy task of keeping track of one target *S*, could consistently fail to accurately count the number of cigarettes smoked. In trying to clarify this puzzling observation, McFall reports an item of data that was unintentionally collected concerning a day the class did not meet. Several of the *S*s monitoring their own smoking behavior—though only one of the *O*s—reported smoking data for that particular day, supporting the belief that smoking *S*s made up their self-report data at some later date, rather than noting down the number of cigarettes smoked each day as they had been instructed to do.

Quite appropriately, McFall points to the need of independent checks on the validity of self-monitoring data and the importance of not confounding self-report and task motivation. It seems, however, that the factors which may determine the data obtained in the reported study may be even more subtle and complex. Thus, when comparing the base-line performance with that during the experimental period, it was shown that there was a significant increase in smoking behavior among those *S*s instructed to report the number of cigarettes smoked and an equally dramatic decrease among those *S*s instructed to report the number of times they felt like smoking but did not.

To understand what appears to have occurred in this study, it is essential to consider the experiment from the student-volunteer participant's point of view. Ignoring for a moment what probably were at least partially unsuccessful deception aspects of the experiment, the student smoker finds himself in a class with a professor interested in smoking behavior. There is a "No Smoking" sign in the room. However, not only does the professor ignore the sign personally, but he also explains that it is relevant only to afternoon classes. Finally, the class is asked to participate in an experiment on smoking, being assigned to one of two groups. He is aware that in addition to *S*s receiving his instructions, there are others receiving other instructions. As I have indicated, smoking is a conspicuous behavior. From the *S*'s point of view, whether he smokes or not is clearly obvious to his instructor, who is also the *E*. While the investigator may feel that the task of keeping track of smoking behavior of his 16 *S*s is superhuman, no such assumption needs to be made by the *S*s. The fact is that the investigator himself could easily have kept track of two or even three *S*s' smoking behavior on any given day without difficulty; in no way, therefore, ought one to assume that *S*s considered their smoking in this class as private behavior. On the contrary, they likely knew the investigator's interest, realized that they were participating in a study important to him, and undoubtedly surmised that he had hypotheses about what they would do.

The actual smoking behavior of the *S*s as reported supports such a conjecture. Thus, whereas self-reported smoking behavior would tend to lead to a decreased base line in most instances, it was found here to have led to a dramatic increase in reported behavior. It is likely that the professor communicated, by his modeling behavior, that he wanted smoking behavior and expected it to be augmented.

Such an intepretation is consonant with the even greater shift in the cigarette consumption of the no-smoke group. Again, it stands to reason that asking *S*s to report whenever they felt like smoking but did not is an implicit request to decrease one's smoking. It is not surprising therefore that *S*s did so dramatically. Unfortunately, these findings are documented only for the public smoking behavior. Public, in this sense, is intended to mean smoking behavior in the sight of and with the knowledge of the professor. Smoking behavior had, of course, been designated as an experimentally relevant variable. Without detailed inquiry data, it is difficult to know much about the *S*s' motives and the reasons why the one group increased their smoking behavior above base line. The possibility that they expected a request to stop smoking at some future time or some kind of crossover design cannot be excluded. Expectational effects of this kind on "base-line performance" have been elegantly demonstrated by Zamansky, Scharf, and Brightbill (1964) in hypnotic research.

A discussion of this kind when translated into experimental terms invariably emphasizes the need for additional control groups. Unfortunately, it is impossible to design an experiment where a colleague cannot think of additional groups that should have been run. The solution does not lie in increasingly complex factorial designs, but rather in a recognition of the limitations inherent in any single study. In this instance I have emphasized the public, nonprivate character of the period when smoking behavior was being recorded, because I believe such public behavior could, for short periods, easily be modified. For example, if a class which included smokers were divided into two groups, half of which were asked by the instructor, "Please, for the sake of an experiment that is important, increase the amount of smoking you do over the next few lectures," and the other half were asked, "Please, for the

sake of an experiment that is important, decrease the amount of smoking you do over the next few periods," one might reasonably anticipate significant effects. Assuming that findings of this kind were easily obtainable, extreme caution would be essential in interpreting the data in McFall's study. In other words, if we are obtaining data under conditions where *S*s can easily alter their performance without much cost to themselves, inference must be drawn exceedingly cautiously. To again use the example of obesity, it is one thing to demonstrate that a given diet results in modification of eating behavior whenever the patient eats with his therapist, but it is quite another to demonstrate weight loss over a prolonged period (though even the latter could, of course, be affected by simple instructions in some instances). In the case of the former, such a finding would have little interest, while in the case of the latter, we would weigh it entirely differently.

There are some situations when problems of compliance are relatively unimportant, such as when the test procedures involve stable maximal performances, as in the case of athletes. Thus, it would be of interest to demonstrate that instructions can shorten the time required for an experienced track man to cover a mile, whereas to show that instructions can lead to an increase in time would not necessarily have much significance. The same is true with studies involving endurance, memory, learning, etc.—all instances where reliable, stable base lines are obtained under motivated conditions. Even then, however, it would seem best to determine the ease with which such a base line could be modified by asking an individual to do so. Whenever a situation exists where *S* can easily alter his performance in a given direction when asked to do so, caution is essential in interpreting similar data obtained from human *S*s, even if the techniques employed appear different.

One way in which some of these problems can be circumvented is to study what the *S* believes to be private behavior. Those measures that are believed to be nonreactive all share the quality that it is intended that *S* does not recognize our interest in them (see Webb, Campbell, Schwartz, & Sechrest, 1966). Under these circumstances, we may reasonably hope that he is unaware of the measure. However, most of these measures are nonreactive only to the extent that *S* fails to recognize our interest. They, therefore, share many of the qualities of deception experiments and require special procedures (which I have called quasi-controls: Orne, 1969) in order to make certain whether *S* did or did not perceive the situation to involve private behavior.

The study of the effects of self-monitoring on normal smoking behavior utilizes a nonrepresentative public sample of smoking behavior in order to draw inference about how much the individual actually smokes. It seems likely to me that the findings were a function of the demand characteristics of the particular experimental situation and therefore do not allow more general inference. Thus, the experimental data need not reflect the *S*s' overall smoking behavior. Intuitively I am convinced the conclusions about potential effects of self-monitoring are correct, while I am equally certain, without the benefit of specific data, that the findings concerning the effects of differential instructions will not generalize beyond the experimental situation.[3]

The kind of demonstration that would be required to validate the effect of instructional interventions in smoking behavior is extremely difficult to obtain. The same is true in many other clinical situations, and probably because of these difficulties, many behavior therapists have chosen to ignore the problems inherent in self-reporting and use this measure as though it were fully valid. To the extent that one is dealing with patients who are paying for treatment, as opposed to experimental *S*s who are being paid, the motivational factors may be given different weights. Not that patients' self-reports are necessarily more accurate, but in the absence of a highly meaningful relationship, the individual who seeks help is more likely to be concerned about what the data mean to him than what they mean to the therapist-experimenter. However, the need for independent evaluation of the findings remains. To the extent that the evaluation procedures are public and often of short duration, the possibility that changes obtained in them may not be representative of the patient's overall behavior must be kept in mind. For example, the many snake phobia studies which follow the Lang and Lazovik (1963) model all use a behavioral snake-avoidance test. The possibility that this test is reactive to the implied wishes of the *E* cannot be excluded (D. A. Bernstein, personal communication, 1969). Thus, Orne and Evans (1965) showed that unselected *S*s could be induced to carry out apparently dangerous and self-destructive actions which neither other *S*s nor colleagues thought likely.

[3] That is, the effect would not be materially different from communicating to the *S* the *E*'s desire that he should increase or decrease his smoking behavior.

There are no simple ways around the problems inherent in criterion measures of change. They will have to be evaluated in each instance. Fortunately, they become less serious when we are dealing with items of behavior that the *S* cannot or will not undertake prior to treatment. The minimum requirement would be to show that a simple request is not effective in eliciting the particular behavior at the onset. Further, when we are dealing with clear-cut end points, *Ss'* reports tend to become more reliable. For example, a statement by a patient that he has decreased the number of cigarettes consumed is likely to be far less reliable than a statement that he has stopped smoking over an extended period of time. The latter report, in order to be inaccurate, would demand that the patient consciously lie, whereas the former statement could easily be subject to self-deception.

In many instances where we are dealing with serious disturbances of functioning, it is possible to obtain evidence about the individual at work, in school, or similar situations where others can corroborate self-reports. Such reports by others are generally less responsive to subtle changes; they also tend to be more reliable. Here, again, reports of quantitatively greater changes are more likely to be trustworthy. It would seem, then, that in order to study the effect of any therapeutic technique, it would be best to use *S* populations where large and unequivocal changes beyond the individual's voluntary capacity at the onset of treatment can be obtained. The use of college student volunteers not only makes inference about treatment tenuous, but also has limited scientific significance due to the general inadequacy of the criterion measures that are employed. The technique of using extreme cases, be they severe behavior pathology or other examples of profound effects, not only has face validity and therapeutic significance, but also may be the most effective way of minimizing those aspects of the experimental situation which interfere with reliable and logically valid findings.

To summarize, McFall's basic observation that self-report measures may in themselves be reactive is likely to be true, and while I do not think that his other conclusions are likely to be generalizable beyond the concrete experimental situation, the study itself raises important issues. Hopefully, recognition of the possible distortions introduced by various criterion measures of change employed in evaluating treatment will lead to more sophisticated and less easily influenced criteria.

REFERENCES

LANG, P. J., & LAZOVIK, A. D. Experimental desensitization of a phobia. *Journal of Abnormal and Social Psychology,* 1963, **66**, 519–525.

MCFALL, R. M. Effects of self-monitoring on normal smoking behavior. *Journal of Consulting and Clinical Psychology,* 1970, **35**, 135–142.

ORNE, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist,* 1962, **17**, 776–783.

ORNE, M. T. Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. Rosnow (Eds.), *Artifact in behavioral research.* New York: Academic Press, 1969.

ORNE, M. T., & EVANS, F. J. Social control in the psychological experiment: Antisocial behavior and hypnosis. *Journal of Personality and Social Psychology,* 1965, **1**, 189–200.

ORNE, M. T., & HOLLAND, C. H. On the ecological validity of laboratory deception. *International Journal of Psychiatry,* 1968, **6**, 282–293.

WEBB, E. J., CAMPBELL, D. T., SCHWARTZ, R. D., & SECHREST, L. *Unobtrusive measures: Nonreactive research in the social sciences.* Chicago: Rand McNally, 1966.

ZAMANSKY, H. S., SCHARF, B., & BRIGHTBILL, R. The effect of expectancy for hypnosis on prehypnotic performance. *Journal of Personality,* 1964, **32**, 236–248.

(Received January 22, 1970)